CORRELACION

Notas

Indice

INDICE 1
1. OBJETIVOS DEL TEMA 1
2. CONCEPTOS BÁSICOS PREVIOS 2
3. INTRODUCCIÓN 2
4. APUNTE HISTÓRICO 3
5. GRÁFICOS DE DISPERSIÓN 4
6. EL COEFICIENTE DE CORRELACIÓN DE PEARSON, r 5
7. CÁLCULO DEL COEFICIENTE DE CORRELACIÓN r 8
8. ESTADÍSTICA INFERENCIAL 10
9. EL COEFICIENTE DE DETERMINACIÓN 11
10. OTRAS MEDIDAS DE CORRELACIÓN. (1) SIMPLIFICACIÓN DEL CÁLCULO DEL COEFICIENTE DE CORRELACIÓN 11
11. OTRAS MEDIDAS DE LA CORRELACIÓN. (2) COEFICIENTE DE CORRELACIÓN NO PARAMÉTRICO 14
12. MEDIDAS DE ASOCIACIÓN 16
13. OTRAS MEDIDAS DE ASOCIACIÓN RELACIONADAS CON EL ANOVA 18
14. COEFICIENTE DE CORRELACIÓN PARCIAL Y SEMIPARCIAL: CONTROL DEL IMPACTO DE OTRAS VARIABLES SOBRE LA CORRELACIÓN 18
15. EXTENSIÓN MULTIVARIADA 19
BIBLIOGRAFÍA 20

1. Objetivos del tema

- Conocer y comprender el concepto de correlación
- Saber cuándo debe utilizarse técnicas de correlación y sus limitaciones
- Saber representar e interpretar un diagrama de dispersión
- Saber leer los resultados de un análisis de correlación producido por un sistema informático
- Saber calcular e interpretar un coeficiente de correlación
- Saber calcular e interpretar los intervalos de confianza del coeficiente de correlación
- Conocer, saber cuándo utilizar y como interpretar las medidas de correlación alternativas al coeficiente de correlación de Pearson
- Saber calcular e interpretar el coeficiente de determinación
- Conocer, saber cuándo utilizar y como interpretar los coeficientes de correlación parcial y semiparcial
- Conocer, saber cuándo utilizar y como interpretar el coeficiente de correlación múltiple

2. Conceptos básicos previos

2.1. Varianza y cuasivarianza

En las notas anteriores se ha descrito la varianza (Descriptivos, página 4) y la cuasivarianza (Muestreo, página 6) como una medida de dispersión de los datos:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2$$

2.2. Covarianza

La covarianza s_{xy}^2 es una medida de cuantificación de la concordancia de la desviación de dos variables:

$$s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

$$s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i\, y_i - \frac{1}{n}\sum_{i=1}^{n}x_i \sum_{i=1}^{n}y_i$$

2.2. Estandarización

La estandarización es un proceso de cálculo para lograr variables comparables en magnitud y dispersión. Consiste en restar a cada observación la media y dividir este resultado por la desviación típica. Se transforma en una distribución de media cero y varianza igual a uno.

3. Introducción

El estudio de la relación entre dos o más variables tiene una gran importancia en el Laboratorio Clínico. Además de las aplicaciones en investigación básica, aplicada y clínica, los problemas de calibración, validación y comparación metodológica son corrientes en el día a día.

En general, cuando se habla de medir relaciones entre dos o más variables, se habla de dos tipos de análisis: regresión y correlación. El análisis de la regresión se utiliza para establecer las relaciones entre dos variables, habitualmente en el contexto de un estudio prospectivo o experimental. El análisis de la correlación, sin embargo, pretende averiguar la intensidad o fortaleza de esta relación, la proximidad de los puntos a la línea (recta) de regresión y acostumbra a hacerse en el contexto de un estudio retrospectivo u observacional.

La correlación es la relación medible matemáticamente mediante un número que representa la intensidad de la relación, pero no la causalidad.

La correlación sirve para:

(a) explorar la existencia de la relación particular estadísticamente significativa entre las dos variables, es decir, si los cambios en una son consistentes en la otra;
(b) conocer si la relación es positiva o negativa;
(c) cuantificar el grado de significación estadística de la relación, es decir, la confianza (estadística) relacionada con dicha relación;
(d) averiguar que parte de la variación de una variable es explicada por la otra.

El análisis de la correlación implica los siguientes pasos:

(a) el estudio descriptivo mediante el “gráfico de dispersión”;
(b) la estimación del coeficiente de correlación (incluyendo su intervalo de confianza);
(c) la valoración de este coeficiente de correlación (signo y magnitud) y la significación estadística;
(d) la interpretación del coeficiente de correlación evaluando el coeficiente de determinación.

El análisis de la correlación tiene unas aplicaciones concretas pero tiende a ser sobreutilizado y malinterpretado, especialmente quizás porque se asocia una alta correlación con causalidad (y viceversa).

El análisis de correlación es preferible reservarlo para generar hipótesis en lugar de para comprobarlas.

4. Apunte histórico

Francis Galton fue el primero en medir la correlación en 1888 a propósito del estudio que reaslizó para probar la “regresión a la mediocridad” de las alturas de los hijos y la de sus padres, mencionando por primera vez un índice de “co-relación”.



Francis Galton (1822-1911)

F. Y. Edgeworth fue el primero en utilizar el término “coeficiente de correlación” en 1892.



Francis Y. Edgeworth (1845-1926)

La fórmula de cálculo del coeficiente de correlación tal como se conoce hoy es debida a K. Pearson que la publicó en dos trabajos aparecidos en 1896. También fue K. Pearson el primero en hacer inferencia con grandes muestras.



Karl Pearson (1857-1936)

Sin embargo el método utilizado para estas inferencias no era demasiado bueno y Student (William Gosset) lo demostró así en 1908.



William S. Gosset (1876-1937)

En 1915, R.A. Fisher confirmó la apreciación de Student, obteniendo además la distribución exacta del coeficiente de correlación. Fisher propuso una transformación aproximadamente normal en 1921 que incluyó más adelante en su influyente obra *Statistical methods for research workers* haciéndola accesible a la comunidad científica. Desde 1932 se conoce como transformación de Fisher.



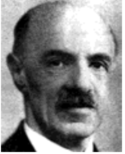
Ronald A. Fisher (1890-1962)

En 1896, G. U. Yule acuñó el término de “coeficientes netos” en oposición a “coeficientes brutos” para los coeficientes de correlación entre dos variables cuando se pretende eliminar el efecto de la variación de una tercera variable. Pearson (1897) juzgó mas oportuno denominarlos “parcial” y “total”. En 1907, una publicación de Yule ya adoptaba esta denominación (1), aunque ya en 1896 G.U. Yule introdujo también el concepto de correlación múltiple. Al año siguiente se refería a él como el “coeficiente de doble correlación R_1”y en 1907 presentó el “enésimo-coeficente de correlación” (*coefficient of n-fold correlation*) al que denominó $R_{1234...}$ y al que K Pearson rebautizó en 1914 (y 1915) como coeficente de “correlación múltiple”.



George U. Yule (1871-1951)

Otras aportaciones notables fueron obra de C. Spearman y M. Kendall. C. Spearman describió en 1904 el coeficiente de correlación ordinal (basado en la posición ordinal de la observación y no en su valor absoluto), y M. Kendall una alternativa diferente al de no paramétrica al de Pearson



Charles Spearman (1863-1945).



Maurice Kendall (1907-1983)

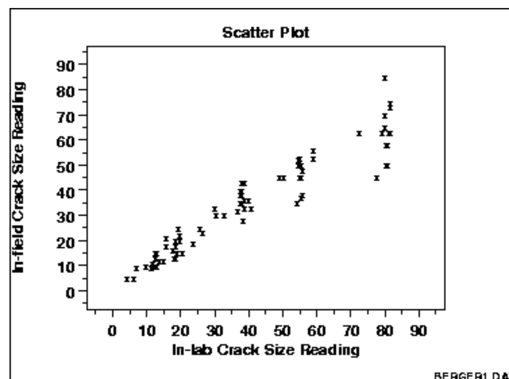
5. Gráficos de dispersión

5.1. Sinónimos

Scatter diagram, *scatter plot*.

5.2. Definición

Es el procedimiento de representación gráfica de las observaciones o mediciones efectuadas en cada individuo[1].



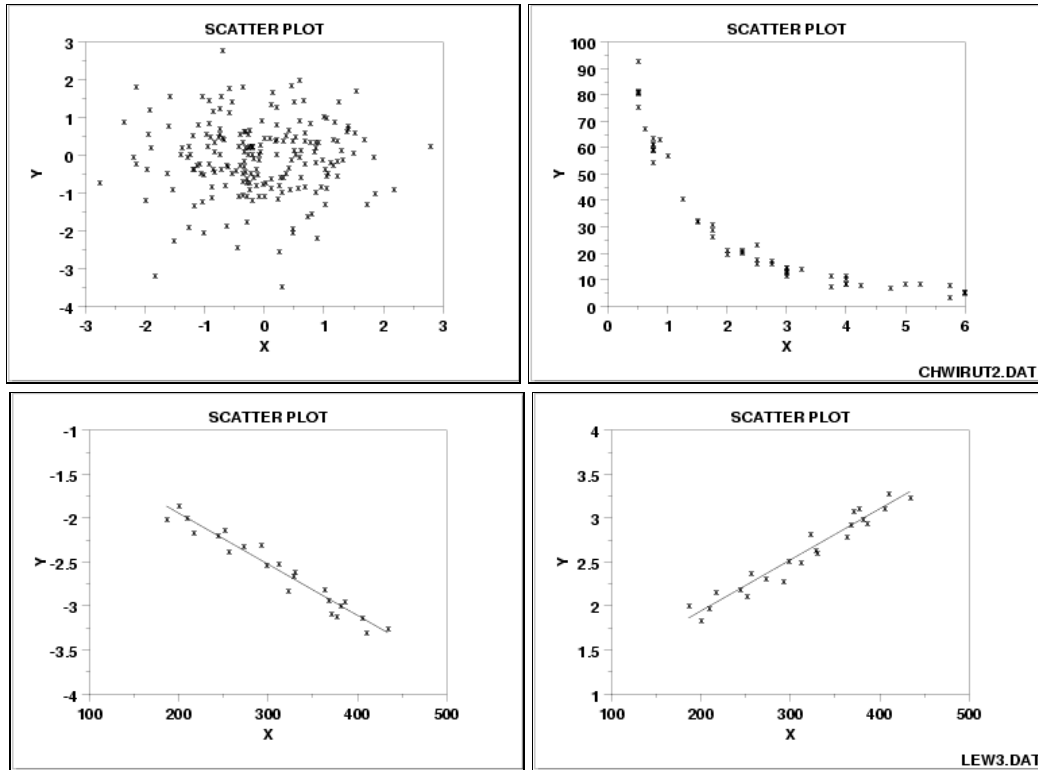
5.3.Utilización del gráfico de dispersión

El gráfico de dispersión se utiliza como una primera aproximación para tener una idea global del tipo de relación existente entre dos variables.

5.4. Interpretación

Cuando existe buena correlación, los puntos se disponen de forma estrecha alrededor de una diagonal del sistema cartesiano: la que tiene su origen en el extremo inferior izquierdo y va al superior derecho si es positiva y al revés si la correlación negativa. Cuando la disposición de los puntos es aleatoria, la correlación es nula.

[1] http://www.ncsu.edu/ncsu/chemistry/resource/excel/excel.html
http://www.ncsu.edu/labwrite/res/gt/graphtut-home.html



5.5. *Limitaciones*

No permite la cuantificación de esta relación, cosa que exige el cálculo del coeficiente de correlación.

6. El coeficiente de correlación de Pearson, r

6.1. *Sinónimos*

Coeficiente r de correlación muestral del producto de momentos.

6.2. *Definición*

El coeficiente de correlación r es la expresión matemática de la relación entre las dos variables aleatorias. Es una versión estandarizada de la covarianza entre X e Y:

$$r = \frac{s_{xy}}{s_x \, s_y}$$

6.3. *Utilización del coeficiente de correlación r*

El coeficiente de correlación r se utiliza para:

(a) comprobar que existe una relación lineal entre dos variables aleatorias, antes de proceder al análisis de regresión;

(b) resumir en un solo número (r) la intensidad de la relación lineal entre estas dos variables.

El coeficiente de correlación r no debe utilizarse para:

(a) establecer relaciones causales entre dos variables;

(b) suplantar el análisis de regresión;

(c) analizar la coherencia entre mediciones.

6.4. Propiedades del coeficiente de correlación r

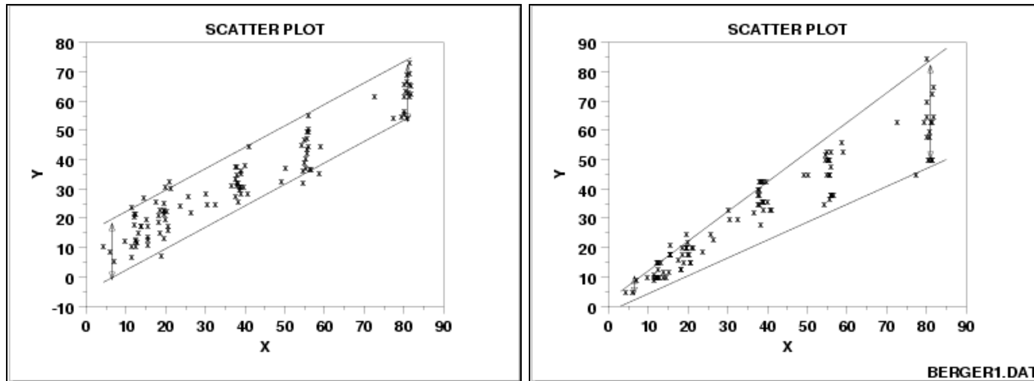
- El coeficiente de correlación r no tiene unidades;
- El intervalo de posibles valores de r es: $-1 \leq r \leq 1$;
- En el análisis de la correlación no es aplicable la distinción entre variable “dependiente” o “independiente” como sucede en el modelo de regresión: aquí ambas variables son “dependientes”.

6.5. Requisitos para calcular el coeficiente de correlación r

- Mínimo de dos observaciones por individuo, medidas en una escala de intervalo (continua); aunque también pueden estarlo en escala ordinal, por codificación de variables categóricas;
- El conjunto de observaciones, la muestra, ha de ser representativo de la población: la presencia de grupos heterogéneos en una muestra (por ejemplo el incluir individuos jóvenes y ancianos en la misma muestra par estudiar la relación entre la edad y el filtrado glomerular) puede resultar inapropiada;
- El conjunto de observaciones no debe contener datos extremos, ya que el coeficiente de correlación r es muy sensible a su presencia;
- Las observaciones no deben contener errores de medición o éste ha de ser relativamente ínfimo ya que la baja fiabilidad de las observaciones atenúa el valor de r ;
- Las variables deben estar relacionadas linealmente, es decir el gráfico de dispersión sigue una diagonal: las relaciones curvilíneas pueden producir r no significativos;
- La forma de la distribución de las variables debe ser igual. Si no tienen la misma distribución, aunque el ajuste sea perfecto, se observará un $r < 1$, y cuanto menos se parezcan las distribuciones más se atenuará r. Este efecto es importante cuando se correlaciona una variable en escala de intervalo con otra ordinal o dicotomizada;
- Además, la distribución de los pares (x_i , y_i) debe ser bivariada normal[2]. Es importante desde el punto de vista inferencial, cuando se trata de valorar la intensidad y la significación estadística de la correlación. Cuando la distribución no sea normal, el intervalo de r puede que no sea $[-1 , 1]$. Sin embargo, el teorema central del límite demuestra que para muestras grandes los índices implicados en las pruebas de significación se distribuyen normalmente incluso cuando las propias variables no lo sean. De cualquier forma, cuando se prefiera evitar este tipo de conflicto, puede recurrirse a utilizar un cálculo no paramétrico como la ρ de Spearman o un estadístico no paramétrico como la τ de Kendall;
- La varianza de las variables debe ser homogénea (variables homoscedásticas) y no restringida. Si la varianza es truncada o restringida en una o varias variables, por ejemplo por un muestreo deficiente, el coeficiente de correlación puede verse afectado. También por truncación del intervalo de la variable por dicotomización de datos continuos o reducción de la escala.

---

[2] El análisis de la r de Pearson es esencialmente paramétrico, basado en la suposición que los pares (x_i , y_i) siguen una distribución normal bivariada. La distribución nomal bivariada es una extensión de la distribución normal caracterizada por que los valores de X e Y se distribuyen normalmente y la distribución marginal de (X , Y) también es normal, con cinco parámetros ($\overline{x} , \overline{y} , \sigma_X^2 , \sigma_Y^2 , \rho$). La evaluación de esta normalidad es ciertamente problemática, aunque una regla aproximada dice que si X e Y se distribuyen normalmente, puede actuarse como si (X , Y) fuera también normal. En cualquier caso de dudas, puede optarse por el recurso de estimar el coeficiente de correlación de forma no paramétrica.



6.6. Interpretación del coeficiente de correlación r

Para interpretar el coeficiente de correlación r hay que conocer las magnitudes en que se han medido las variables.

El coeficiente de correlación r mide sólo la relación lineal entre variables, es decir, es una medida del poder explicativo del modelo lineal, pero no mide la magnitud de la pendiente de la regresión ni si es adecuado un modelo lineal. Si la relación sigue un modelo no lineal (curvado) el coeficiente de correlación puede ser 0. En otras palabras, un coeficiente de correlación pequeño no indica necesariamente que hay poca relación entre las variables, porque puede estar reducido si la relación entre estas variables no es lineal. Una medida de precaución es calcular el intervalo de confianza del coeficiente de correlación r para tener una visión más general, especialmente cuando el número de observaciones sea pequeño.

El coeficiente de correlación r no implica causalidad. Para establecer una relación causa efecto entre dos variables es preciso que la variable causal preceda a la variable causada y cumplir una serie de condiciones (reversibilidad, fortaleza, respuesta a la exposición) contenidas en el llamado modelo de Rubin/Bradford Hill.

Desde el punto de vista geométrico, los conjuntos de n observaciones son vectores en un espacio k–dimensional (habitualmente $k=2$, ya que si $k>2$ se hablará de correlación múltiple) y el coeficiente de correlación r sería el coseno del ángulo entre los vectores centrados por sustracción de la respectiva media. Si estos vectores son perpendiculares entre sí (es decir forman un ángulo recto) porque se trata de variables no correlacionadas, será $r=0$ y las variables se denominarán ortogonales (del griego “ángulo recto”)

El coeficiente de correlación r siempre sobrestima la relación lineal entre dos variables en la población, por lo que requiere una corrección para conseguir una mejor estimación. El resultado de esta corrección se denomina “coeficiente de correlación ajustado” y se representará como $\tilde{r}$.

6.6.1. Signo

El signo de r es positivo cuando al aumentar o disminuir el valor de una variable aleatoria la otra aumenta o disminuye simultáneamente.

El signo de r es negativo cuando al aumentar o disminuir el valor de una variable aleatoria la otra disminuye o aumenta simultáneamente.

6.6.2. Magnitud

Si las variables X e Y son independientes, $r=0$. Sin embargo, la inversa no es siempre cierta, salvo que ambas variables se distribuyan normalmente;

Si las variables X e Y están perfectamente relacionadas entre sí:

$$|r|=1$$

6.7. Limitaciones del coeficiente de correlación r

6.7.1. Causalidad

La principal limitación que tiene el coeficiente de correlación es la posibilidad de mala interpretación como demostración de causalidad. Cuando dos variables tienen un coeficiente de correlación altamente

significativo no implica en modo alguno que una sea causa de la otra. Puede ser, por ejemplo, que esta alta correlación se deba a la existencia de una tercera (o más) variables dependientes implicadas en el proceso. En estos casos se recomienda recurrir al cálculo del coeficiente de correlación parcial o semiparcial.

En ocasiones se califica de "espúrea" la alta correlación no explicable.

6.7.2. Significación paradójica

Una segunda limitación se origina en el caso de estudiar la correlación en una muestra cuyo contingente n sea muy grande. En estos casos, r tiende a ser significativo aunque su valor absoluto no sea demasiado grande. Cuando se evalúa el coeficiente de determinación R^2 puede darse la circunstancia de que a pesar de la alta significación estadística de r el coeficiente R^2 exprese sólo un insignificante porcentaje de variación de una variable es explicado por la segunda variable. Esta paradoja resume el escaso valor que puede tener el coeficiente r para definir la relación entre variables.

6.7.2. Evaluación de medidas de error

El coeficiente de correlación no debe ser utilizado en comparación de métodos: indica el grado de asociación, no el nivel de acuerdo o concordancia entre diferentes métodos.

7. Cálculo del coeficiente de correlación r

7.1. Estandarización de X e Y

Una forma es utilizar el producto cruzado de los valores estandarizados de las variables:

$$z_x = \frac{x - \overline{x}}{s_x}$$

$$z_y = \frac{y - \overline{y}}{s_y}$$

$$r = \frac{\sum_i \left| z_{x,i}\, z_{y,i} \right|}{n} = \frac{\sum_i \frac{x_i - \overline{x}}{s_x} \sum_i \frac{y_i - \overline{y}}{s_y}}{n} = \frac{1}{n} \sum_i \frac{(x - \overline{x})(y - \overline{y})}{s_x\, s_y}$$

O de forma más directa, mediante:

$$r = \frac{n \sum_i x_i\, y_i - \sum_i \overline{x} \sum_i \overline{y}}{\sqrt{\left[n \sum_i x_i^2 - \left(\sum_i \overline{x} \right)^2 \right] \left[n \sum_i y_i^2 - \left(\sum_i \overline{y} \right)^2 \right]}}$$

Sin embargo, aunque útil para programar informáticamente, esta fórmula puede introducir errores de redondeo cuando se aplica manualmente, recomendándose en esta circunstancia utilizar la fórmula anterior.

7.2. Estimación insesgada

Desde 1915 en que R. A. Fisher lo evidenciara, se sabe que el coeficiente de correlación r de Pearson estimado según las fórmulas anteriores tiende a infraestimar la correlación poblacional. Este sesgo disminuye cuando aumenta el número de datos n, de modo que para contingentes numerosos, el sesgo es mínimo. Fisher propuso como estimador insesgado:

$$r' = r \left(1 + \frac{1 - r^2}{2n} \right)$$

Olkin y Pratt (2) recomendaron utilizar el siguiente estadístico que parece ser que corrige mejor el sesgo de r cuando el efectivo es $n \leq 20$:

$$r' = r\left(1 + \frac{1-r^2}{2n-6}\right)$$

El sesgo del estimador del coeficiente de correlación r de Pearson no debe confundirse con la inestabilidad de la varianza de la que adolece también este estadístico. Fue también R. A. Fisher quien lo evidenció y el que introdujo una transformación (no lineal) denominada z:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

El error estándar de z es:

$$es(z) = \sqrt{\frac{1}{n-3}}$$

aunque Fieller y Pearson prefieren una fórmula más ajustada:

$$es(z) = \sqrt{\frac{1,060}{n-3}}$$

La función inversa (denominada función tangente hiperbólica) es:

$$r_z = \frac{e^{2z}-1}{e^{2z}+1}$$

7.4. Corrección por atenuación

La fiabilidad puede interpretarse como la correlación de una variable consigo misma. La corrección para la atenuación de una correlación r_{xy} es una función de la fiabilidad de dos variables r_{xx} y r_{yy}:

$$r'_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}\, r_{yy}}}$$

7.5. Coeficiente de correlación ajustado $\tilde{r}$

El coeficiente de correlación es un estadístico descriptivo que siempre sobrestima la correlación de la población. Para una estimación más exacta, puede ajustarse según la expresión:

$$\tilde{r} = \sqrt{1-\left|\left(1-r^2\right)\frac{n-1}{n-2}\right|}$$

7.6. Relaciones de r con otros parámetros afines:

(a) con el coeficiente de regresión b (pendiente):

$$r^2 = b\,b'$$

(b) con el coeficiente de determinación R^2:

$$R^2 = r^2$$

(c) con la covarianza s_{xy}^2:

$$s_{xy}^2 = r\, s_x\, s_y$$

7.7. Promedio de coeficientes de correlación

En algunas circunstancias puede ser interesante considerar un valor medio de diferentes coeficientes de correlación. Por ejemplo, si se desea evaluar el rendimiento interensayo en un estudio de aprendizaje, una circunstancia rara en medicina de laboratorio. Más interesante puede ser en un metanálisis. El coeficiente de correlación medio será:

$$\overline{r} = \frac{\sum_i n_i r_i}{\sum_i n_i}$$

donde: n_i es el tamaño muestral de las i muestras;

r_i es el coeficiente de correlación de las i muestras.

Como que r es un estadístico sesgado que para valores superiores a 0 infraestima la correlación poblacional, también lo será $\overline{r}$ .Por la misma razón, al ser r un estadístico de varianza inestable, también lo será $\overline{r}$ y también le será aplicable la transformación z de Fisher para estabilizarla, aunque este último punto es objeto de controversia.

8. Estadística inferencial

(Véase Módulo 2)

8.1. Intervalos de confianza del coeficiente de correlación

Para calcular el intervalo de confianza del coeficiente de correlación se recurre a la transformación de Fisher mencionada anteriormente, que se distribuye normalmente y cuyo error estándar también se mencionó. El intervalo estará comprendido entre los límites inferior z_L y superior z_U :

$$z_L = z - 1{,}96\sqrt{\frac{1}{n-3}}$$

$$z_U = z + 1{,}96\sqrt{\frac{1}{n-3}}$$

Seguidamente, para conseguir estos límites en términos de coeficiente de correlación, se aplicará la transformación inversa:

$$r_L = \frac{e^{2z_L}-1}{e^{2z_L}+1}$$

$$r_U = \frac{e^{2z_U}-1}{e^{2z_U}+1}$$

8.2. Test de hipótesis

Definición del test:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Distribución de probabilidad de r :

$$r \sim N\left(\rho, \frac{1-\rho^2}{n-2}\right)$$

Estadístico a calcular:

$$t_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Alternativamente, ya que $F = t^2$:

$$F_{obs} = \frac{r^2}{1-r^2}(n-2)$$

Cuando la hipótesis nula es cierta, y cumpliéndose los requisitos y suposiciones, los estadísticos observados se distribuyen según t_{n-2} y $F_{1,n-2}$ respectivamente:

$$p = 2\ \Pr\left(t_{n-2} \geq |t_{obs}|\right)$$
$$p = 2\ \Pr\left(F_{1,n-2} \geq |F_{obs}|\right)$$

9. El coeficiente de determinación

El coeficiente de determinación R^2 indica la proporción de varianza común entre una variable dependiente y una(s) variable(s) independientes. Se define como el cuadrado del coeficiente de correlación r de Pearson:

$$R^2 = r^2$$

Sus valores están, pues incluidos en el intervalo $[0,1]$.

En las notas referentes a la Regresión, se trata más ampliamente, en el contexto de la regresión lineal.

10. Otras medidas de correlación. (1) Simplificación del cálculo del coeficiente de correlación

Versiones simplificadas del cálculo del coeficiente de Pearson, a veces denominados “atajos”, son:

- El coeficiente de correlación ρ (rho) de Spearman
- El coeficiente de correlación biserial puntual r_{bp}
- El coeficiente de correlación Φ (phi)

escala	intervalo	ordinal	nominal	dicotómico natural	dicotómico artificial
intervalo	r de Pearson				
ordinal	biserial r_b	ρ de Spearman τ de Kendall			
nominal	biserial puntual r_{bp}	r_{br} biserial ordenado	C de contingencia γ, λ		
dicotómico natural	biserial puntual r_{bp}	r_{bp} biserial puntual	V de Cramer	Φ	
dicotómico artificial	biserial r_b				tetracórico r_{tet}

10.1. El coeficiente de correlación ρ de Spearman

Definición

Es un caso especial de coeficiente de correlación de Pearson en que los datos continuos son previamente convertidos en ordinales.

Propiedades

- Aplicable cuando se trata de datos ordenados (o continuos reducidos a ordinales);
- Menos influenciable por la presencia de datos sesgados o variables;
- Es una medida de correlación no paramétrica en el sentido de que no implica suposiciones acerca de la forma de la distribución de frecuencias de las variables implicadas;
- Tampoco requiere la suposición de linealidad.

Cálculo[3]

Una vez convertidos los datos en sus respectivos ordinales, se aplica el mismo procedimiento de cálculo que para el coeficiente de Pearson.
Una fórmula alternativa, más práctica, es:

$$\rho = 1 - \frac{6\sum d^2}{n(n-1)}$$

donde: d^2 es la diferencia de los ordinales de X e Y;
n es el número de pares de datos

10.2. El coeficiente de correlación biserial puntual

Definición

Es un caso especial de coeficiente de correlación de Pearson en que una variable está medida en escala nominal dicotómica natural (es decir, los datos se expresan como dos alternativas, por ejemplo "1" y "0") y la otra en escala de intervalo o continua.

Propiedades

Aplicable cuando se trata de datos ordenados (o continuos reducidos a ordinales);

Cálculo

$$r_{bp} = \frac{(\bar{y}_1 - \bar{y}_0)\sqrt{pq}}{\sigma_y}$$

donde: $\bar{y}_0, \bar{y}_1$ son las medias de los ordinales de los pares de datos con puntuaciones "1" y "0";
p es la proporción de pares de datos con puntuación "0";
$q = 1 - p$ es la proporción de pares de datos con puntuación "1";
σ_y es la desviación típica de los datos de la población Y.

Interpretación

Como el caso del coeficiente de correlación de Pearson, r_{bp} tiene un intervalo de valores $-1 \le r \le 1$.

[3] Una calculadora "on line" puede encontarse en http://www.wessa.net/rankcorr.wasp

10.3. El coeficiente de correlación biserial r_b

Definición

El coeficiente de correlación biserial r_b es similar al coeficiente de correlación biserial puntual r_{bp} pero para variables medidas en escala cuantitativa una y dicotomizada la otra, aunque subyacentemente se distribuya normalmente.

Cálculo[4]

$$r_b = \frac{\left(\bar{y}_1 - \bar{y}_0\right) \frac{pq}{Y}}{\sigma_y}$$

donde: $\bar{y}_0, \bar{y}_1$ son las medias de las variables cuantitativas;

p es la proporción de pares de datos con puntuación "0";

$q = 1 - p$ es la proporción de pares de datos con puntuación "1";

Y es la altura de la distribución normal estandarizada (media = 0 y varianza = 1) en el punto z tal que

$$\Pr\left(z' < z\right) = p$$
$$\Pr\left(z' > z\right) = q$$

σ_y es la desviación típica de los datos de la población Y.

Interpretación

Dado que el factor pq/Y del numerador siempre será > 1, el coeficiente de correlación biserial r_b será siempre mayor que el coeficiente de correlación biserial puntual r_{bp}.

10.4. El coeficiente de correlación biserial ordinal r_{bo}

Definición

El coeficiente de correlación biserial ordinal r_{bo} es similar al coeficiente de correlación biserial puntual r_{bp} pero para variables medidas en escala ordinal una y dicotomizada la otra, aunque subyacentemente se distribuya normalmente.

Cálculo

$$r_{bo} = 2\frac{\left(\bar{y}_1 - \bar{y}_0\right)}{n}$$

donde: $\bar{y}_0, \bar{y}_1$ son las medias de los ordinales para los pares de datos con puntuaciones "1" y "0";

n es el número de pares de datos.

Interpretación

Salvo por el matiz de la variable dicotomizada, tiene la misma interpretación que el coeficiente de correlación biserial puntual r_{bp}.

---

[4] Una aplicación para Excel puede obtenerse en: http://psych.colorado.edu/~bonk/IF_pbi_template.xls.zip.

10.5. Coeficiente de correlación Φ

Definición

Esta simplificación del cálculo del coeficiente de correlación r de Pearson se ha utilizado en la comprobación de la hipótesis de independencia, es decir de la propiedad inversa a la asociación, cuando ambas variables son dicotómicas. En este caso se acostumbra a establecer unas tablas bidimensionales (una variable en cada dimensión) que representan la frecuencia de cada una de las cuatro categorías y las sumas marginales:

		X		
		+	–	
Y	+	a	b	$a+b$
	–	c	d	$c+d$
		$a+c$	$b+d$	n

Cálculo[5]

$$\Phi = \frac{b\,c - a\,d}{\sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)}}$$

Interpretación

Como el caso del coeficiente de correlación de Pearson, r_{bp} tiene un intervalo de valores $-1 \leq r \leq 1$. Sin embargo, los valores extremos $|r| = 1$ sólo son posibles cuando la suma de las dos filas y la suma de las dos columnas sea igual.

11. Otras medidas de la correlación. (2) Coeficiente de correlación no paramétrico

Cuando se violan los supuestos del cálculo del coeficiente de correlación de Pearson, puede recurrirse a la transformación de estos datos para corregir la anomalía o utilizar un análisis no paramétrico de la correlación.

11.1. Coeficiente de correlación τ de Kendall

Propiedades

- El coeficiente de correlación τ de Kendall tiene un intervalo de valores $-1 \leq \tau \leq 1$.
- Cuando la concordancia o discordancia entre los ordinales de dos variables es perfecta, τ vale 1 o –1; cuando son independientes, $\tau = 0$.

Cálculo[6]

$$\tau = \frac{4\;P}{n(n-1)} - 1$$

Cuando $n > 40$:

[5] Puede obtenerse una calculadora "on line" en http://faculty.vassar.edu/lowry/tab2x2.html

[6] Puede obtenerse una calculadora "on line" en http://www.wessa.net/rwasp_kendall.wasp

$$z = \frac{P}{\sqrt{\frac{n\,(n-1)\,(2n+5)}{18}}}$$

donde: n es el número de pares de observaciones;

P es el número de pares concordantes menos el número de pares discordantes;

z corresponde a la distribución normal.

Interpretación

Cuanto mayor sea el número de pares concordantes, mayor será P y por tanto τ, indicando que los ordinales son consistentes[7].

11.2. Coeficiente de correlación tetracórico r_{tet}

Definición

Es una estimación de r a partir de variables dicotómicas, como φ, que representan distribuciones de valores en escala de intervalo y distribuidas normalmente, pero que por la razón que fuere han sido dicotomizadas.

---

[7] Nota: Los empates no son concordantes ni discordantes. Si hay muchos empates debe considerarse el utilizar fórmulas alternativas (Todas estos coeficientes presentan valores en el intervalo $[-1,1]$ alcanzando los valores extremos en caso de concordancia o discordancia perfecta, y de 0 en caso de ausencia de asociación):

τ_b *de Kendall*. Este estadístico incluye los empates E_X y E_Y. Se define como:

$$\tau_b = \frac{C - D}{\sqrt{(C+D+E_X)\,(C+D+E_Y)}}$$

τ_c *de Kendall*. Este estadístico incluye el menor número de casos no empatados en la variable X y en la Y. Se define como:

$$\tau_c = \frac{2\,B\,(C-D)}{n^2\,(B-1)}$$

d de Somers: Este estadístico considera tanto variables simétricas (en este caso coincide con la τ_b) como dependientes. En este caso, si la variable dependiente es Y:

$$d = \frac{C - D}{C + D + E_Y}$$

donde: B es el menor número de casos no empatados en la variable X y en la Y;

$C = \frac{n\,(n-1)}{2}$ es el número total de concordancias, incluyendo los E empates;

D es el número total de discordancias;

E_X, E_Y es el número de empates en la variable X y en la Y respectivamente; y

E es el número total de empates.

Todas estos coeficientes presentan valores en el intervalo $[-1,1]$ alcanzando los valores extremos n caso de concordancia o discordancia perfecta, y de 0 en caso de ausencia de asociación.

Cálculo[8]

$$r_{tet} = \cos\left(\frac{180}{1+\sqrt{\frac{b\,c}{a\,d}}}\right)$$

Aplicaciones

Se utiliza en el modelado de ecuaciones estructurales (SEM) y por lo general su utilización como medida de asociación se desaconseja.

Según el procedimiento utilizado para calcularlo, el resultado puede ser diferente.

12. Medidas de asociación

Existe una serie de medidas ocasionalmente denominadas "coeficientes de correlación", pero que con más propiedad deben llamarse "de asociación", son parecidas al coeficiente φ.

12.1. Coeficiente de contingencia C de Pearson

Definición

El coeficiente de contingencia C de Pearson es una medida de asociación entre dos variables medidas en escala nominal. Se trata de adaptar el estadístico χ^2 a tablas de dimensiones superiores a 2 x 2. De hecho existe la recomendación de utilizarlo sólo para tablas de 5 x 5 o superiores, ya que por debajo de esta dimensión se infraestima la asociación de las variables.

Cálculo

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Existe una variante (denominada $C*$ de Sakoda) para ajustar C:

$$C* = \frac{C_{observada}}{C_{máxima}} = \frac{C_{observada}}{\sqrt{\frac{\min(filas, columnas) - 1}{\min(filas, columnas)}}}$$

que varía entre $(0,1)$ independientemente de las dimensiones de la tabla

Interpretación

La interpretación tanto de C como de $C*$ no es intuitiva. Pearson lo consideraba una aproximación al coeficiente r ,y puede ser considerado como un porcentaje de su máxima variación posible. Debe tenerse en cuenta que cuanto menor sean las dimensiones de la tabla, más se alejará C de 1(cosa que no sucede con $C*$). Para tablas de 2 x 2 el valor máximo de C es de 0,71.

12.2. Coeficiente V de Cramér

Definición

El coeficiente V descrito por Harald Cramér (3) es una medida de asociación entre variables medidas en escala nominal.

---

[8] Una calculadora "on line" puede encontrarse en: http://www.mhri.edu.au/biostats/DAG%5FStat/.

Cálculo[9]

Utiliza el estadístico χ^2:

$$V = \sqrt{\frac{\chi^2}{n \min(R-1, C-1)}}$$

donde: C es el número de columnas;
R es el número de filas; y
n es el número de observaciones.

En el caso sencillo de tablas 2 x 2, equivale al coeficiente Φ mencionado anteriormente:

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

Interpretación

El estadístico V está dentro del intervalo $[0,1]$. Un valor $V = 0$ significa una concordancia inexistente y $V = 1$ implica una concordancia perfecta.

12.3. Coeficiente λ de Goodman y Kruskal

Definición

El coeficiente λ de Goodman y Kruskal es una medida de asociación entre dos variables medidas en escala nominal, como el coeficiente de contingencia C de Pearson o el coeficiente V de Cramer, pero en este caso no se utiliza el estadístico χ^2.

Cálculo[10]

$$\lambda = \frac{\Pr(error) - \Pr(error|x(t-1))}{\Pr(error)}$$

Ya que las verdaderas probabilidades no se conocen, λ es estimado mediante:

$$\hat{\lambda} = \frac{\sum_{i=1}^{r} \max(n_i) - \max(C_j)}{N - \max(C_j)}$$

donde: N es el número total de observaciones;
$\max(n_i)$ es la mayor frecuencia en la *i*-èsima fila de las r que puede contener la tabla; y
$\max(C_j)$ es el mayor total de las columnas de la tabla.

Este estadístico se distribuye de forma gaussiana con varianza:

$$\mathrm{var}(\hat{\lambda}) = \frac{\left(N - \sum_{i=1}^{r} \max(n_i)\right)\left(\sum_{i=1}^{r} \max(n_i) + \max(C_j) - 2\sum_{i=1}^{r}{}^{*} \max(n_i)\right)}{\left(N - \max(C_j)\right)^3}$$

[9] Una calculadora "on line" puede encontrarse en: http://faculty.vassar.edu/lowry/newcs.html.
[10] Una calculadora "on line" puede encontrarse en: http://www.mhri.edu.au/biostats/DAG%5FStat/.

donde: $\sum_{i=1}^{r} {}^{*}\max(n_i)$ es la suma de las frecuencias máximas de la columna asociada a $\max(C_j)$.

Esto permite introducir un test de hipótesis, calculando el valor z (4):

$$z = \frac{\hat{\lambda} - \lambda_0}{\sqrt{\mathrm{var}(\hat{\lambda})}}$$

Interpretación

El estadístico λ está en el intervalo $[0,1]$. Un valor $\lambda = 0$ significa que el antecedente no ayuda a predecir el consecuente, y $\lambda = 1$ implica una predicción perfecta.

13. Otras medidas de asociación relacionadas con el ANOVA

13.1. Coeficiente de correlación nolineal η (eta)

Definición

El coeficiente de correlación lineal η (eta) (en ocasiones denominado E^2) permite medir la asociación cuando no se cumple el requisito de linealidad. En otras palabras, es una medida de la linealidad de la relación.

Cálculo

Es el coeficiente entre la suma de cuadrados entre grupos (que mide el efecto de la variable agrupadora, es decir, la diferencia entre las medias de los grupos) y la suma de cuadrados total en el ANOVA

Interpretación

Puede interpretarse como el porcentaje de la varianza total de la variable dependiente explicada por la varianza entre categorías (grupos) constituidos por las variables independientes. Vendría a ser un análogo del coeficiente R^2. Cuando la relación no sea lineal, $\eta > R^2$. Es decir, el intervalo de η se extiende más allá del coeficiente de correlación r.

13.2. Coeficiente de correlación intraclase

Es un tipo de medida de correlación basada en el ANOVA

Mide la homogeneidad relativa de varios grupos en razón de la variación total. Por ejemplo se utiliza para valorar la fiabilidad entre clasificadores:

$$r_{ic} = \frac{MS\ entre\ grupos - MS\ intra\ grupos}{MS\ entre\ grupos - (n-1)\ MS\ intra\ grupos}$$

donde: n es el número medio de casos en cada categoría de la variable independiente X.

Es grande y positiva cuando no hay variación entre grupos pero dispersión entre medias.

El intervalo de valores es: $-\frac{1}{n-1} \le r_{ic} \le 1$.

14. Coeficiente de correlación parcial y semiparcial: control del impacto de otras variables sobre la correlación

Cuando se sospecha que una determinada variable Z (denominada de control o de confusión) pueda interferir en la valoración de la relación existente entre dos variables X e Y puede intervenirse:

(a) refinando el diseño experimental: aleatorizando los individuos, limitándose a un solo valor, ajustando los individuos; o

(b) incorporando la variable Z al análisis y calculado el llamado coeficiente de correlación parcial o el coeficiente de correlación semiparcial.

14.1. El coeficiente de correlación parcial

El coeficiente de correlación parcial (de primer orden) es la relación entre dos variables cuando se ha eliminado de cada una de ellas el efecto que sobre ellas tiene una tercera variable. Es decir, tras controlar estadísticamente la influencia de una tercera variable.

Nomenclatura

$r_{XY \cdot Z}$

donde: X , Y son las variables objeto del estudio;

Z es la variable de control

Cálculo

Consiste en estudiar las correlaciones r_{XY}, r_{XZ} , r_{YZ} y combinarlas:

$$r_{XZ.Y} = \frac{r_{XZ} - (r_{XY})(r_{YZ})}{\sqrt{1 - r_{XY}^2}\,\sqrt{1 - r_{XZ}^2}}$$

14.2. El coeficiente de correlación semiparcial

El coeficiente de correlación semiparcial es la correlación existente entre dos variables cuando se ha eliminado de cada una de ellas el efecto que una tercera variable tiene sobre ella.

Nomenclatura

$r_{X(Y \cdot Z)}$ representa la correlación entre las variables X e Y objeto del estudio, eliminando el efecto de la variable de control Z sobre la variable Y.

$r_{Y(X \cdot Z)}$ representa la correlación entre las variables X e Y eliminando el efecto de la variable Z sobre la variable X.

donde: X , Y son las variables objeto del estudio;

Z es la variable de control

Cálculo

Consiste en estudiar las correlaciones r_{XY}, r_{XZ} , r_{YZ} y combinarlas:

$$r_{X(Y \cdot Z)} = \frac{r_{XY} - r_{XZ}\, r_{YZ}}{\sqrt{1 - r_{YZ}^2}}$$

$$r_{Y(X \cdot Z)} = \frac{r_{XY} - r_{XZ}\, r_{YZ}}{\sqrt{1 - r_{XZ}^2}}$$

15. Extensión multivariada

(Véase el Módulo 3: “Correlación múltiple”).

Bibliografía

1. Yule GU. On the theory of correlation for any number of variables, treated by a new system of notation. Proc R Soc, Ser A 1907;79:182-193.

2. Olkin I, Prat JW. Unbiased estimation of certain correlation coeffcients. Ann Math Stat 1958;29:201-211.

3. Cramér H. Mathematical Methods of Statistics. Princeton: University Press; 1999.

4. Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences. 2ª ed. ed. New York: McGraw-Hill; 1988.